



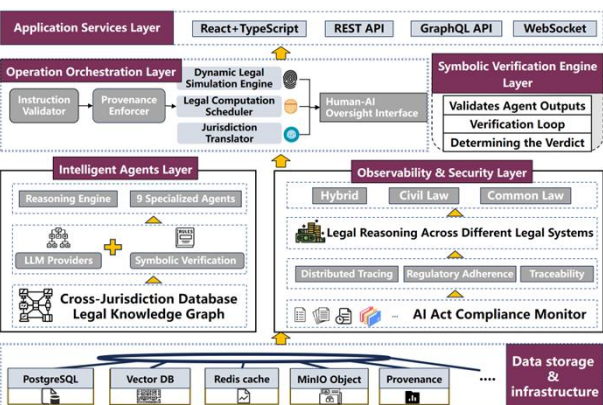
Legal-ISA: A Modular Framework for Systematic Legal AI Evaluation

Daisong Gong
Nankai University

Introduction

Current legal AI research is constrained by **horizontal fragmentation** (narrow domain specialization lacking holistic reasoning) and **vertical fragmentation** (jurisdictional confinement preventing cross-border application), as evidenced by recent benchmarks like LEXam and MSLR. To address this, we introduce **Legal-ISA**, an integration layer inspired by computer architecture. By defining standardized operation interfaces for retrieval and reasoning, Legal-ISA ensures component substitutability and enables systematic, unified evaluation across diverse legal domains and jurisdictions.

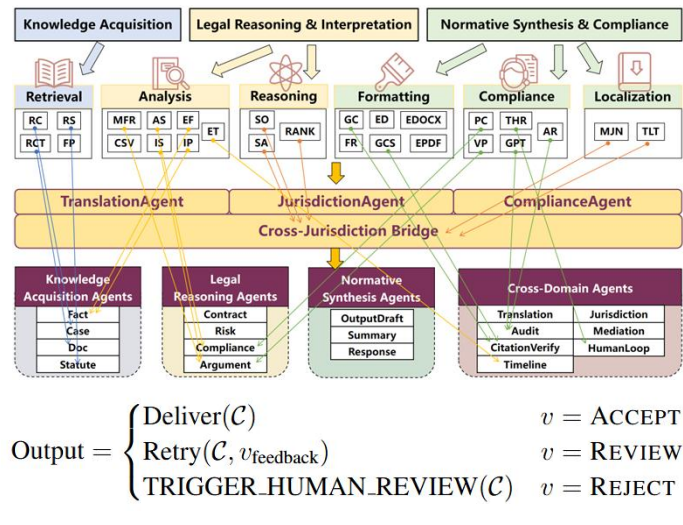
Architecture



$$\text{conf}_{\text{final}} = \text{conf}_{\text{LLM}} \times \left(1 - \beta \times \frac{\#_{\text{unattrb}}}{\#_{\text{total}}}\right)$$

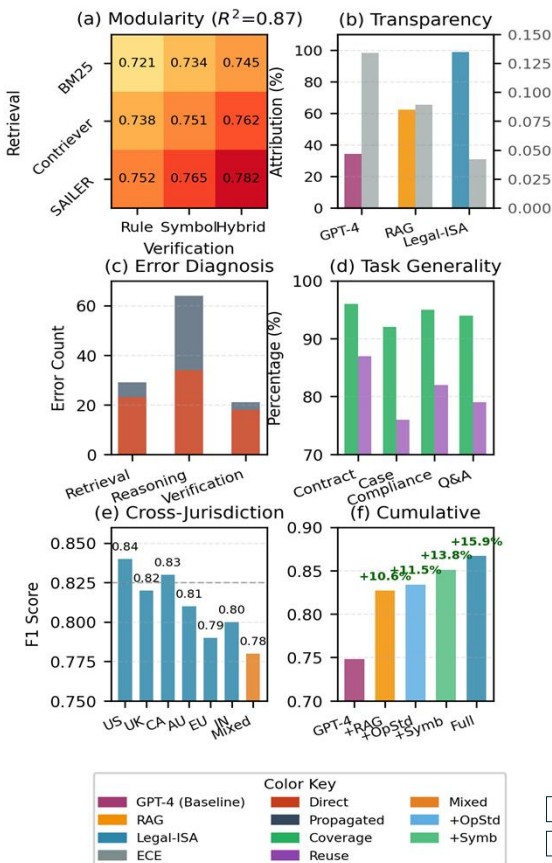
- Architecture:** Six-layer framework integrates reasoning techniques via standardized interfaces and mandatory provenance protocols.
- Orchestration:** Orchestration layer manages validation, scheduling, cross-jurisdictional translation, and human oversight escalation.
- Verification:** Agents execute via a mandatory neuro-symbolic loop correcting outputs against formal legal constraints.
- Assurance:** Assurance layer guarantees observability and auditability backed by polyglot persistence and immutable ledgers.

Mapping



Output = $\begin{cases} \text{Deliver}(C) & v = \text{ACCEPT} \\ \text{Retry}(C, v_{\text{feedback}}) & v = \text{REVIEW} \\ \text{TRIGGER_HUMAN_REVIEW}(C) & v = \text{REJECT} \end{cases}$

Experiment



- System Reliability:** Validates high modular attribution and superior transparency with minimized calibration error compared to baselines.
 - Diagnostic Precision:** Localizes primary errors to reasoning bottlenecks (89% propagation) while maintaining 94.3% functional coverage across diverse legal tasks.
 - Global Impact:** Demonstrates robust cross-jurisdictional performance and achieves a cumulative +15.9% improvement over standard models.
- $R^2 = 0.87$
 $\rho = 0.94$

Category	Configuration	QLAR		LegalBench		CAIL		MultiLegalPile	
		R@10	MRR	F1	Prec	Acc	F1	F1	Prec
Pure Neural Baselines									
	GPT-4o	0.521	0.342	0.748	0.731	0.693	0.701	0.652	0.638
	Claude 3.7 Sonnet	0.536	0.357	0.762	0.745	0.708	0.715	0.668	0.653
	DeepSeek-V3	0.529	0.351	0.756	0.738	0.705	0.712	0.661	0.647
Retrieval-Augmented Generation									
	GPT-4o + BM25	0.638	0.445	0.794	0.781	0.742	0.749	0.718	0.704
	GPT-4o + Contriever	0.662	0.468	0.812	0.798	0.758	0.765	0.731	0.717
	GPT-4o + SAILER	0.681	0.487	0.827	0.813	0.771	0.778	0.745	0.731
	GPT-4o + ColBERTv2	0.677	0.482	0.823	0.809	0.768	0.775	0.741	0.727
Framework Ablations									
	w/o OpStd	0.689	0.493	0.834	0.820	0.779	0.786	0.751	0.737
	w/o Symbolic	0.703	0.507	0.851	0.837	0.794	0.801	0.768	0.754
	w/o Hybrid	0.691	0.495	0.838	0.824	0.783	0.790	0.754	0.740
Legal-ISA (Full)									
	OpStd + Symbolic + Hybrid	0.724	0.521	0.867	0.853	0.812	0.819	0.782	0.768
Improvement over best RAG		+6.3%	+7.0%	+4.8%	+4.9%	+5.3%	+5.3%	+5.0%	+5.1%

Conclusion

- Legal-ISA delivers high transparency and calibrated confidence, enabling principled human oversight and reasoning across diverse legal tasks.
- The framework supports cross-jurisdictional reasoning through human-in-the-loop knowledge engineering, opening collaborative research avenues.
- Our current work explores a legal-agnostic intermediate representation to unify cross-jurisdictional disparities and advance a global legal reasoning.
- We welcome discussions with fellow researchers—feel free to reach us at 2312325@mail.nankai.edu.cn!